SASR-Eval v1: Quantitative Evaluation of Commercial and Fine-Tuned ASR Systems on Somali Speech with Ground-Truth References

SASR-Eval v1 Report

June 22, 2026

Abstract

This report presents SASR-Eval v1, the first ground-truth evaluation of commercial and fine-tuned ASR systems on Somali speech. Using 50 stitched clips (∼10 seconds each, ∼550 seconds total) drawn from the `skydheere/soomali-asr-dataset` HuggingFace test split, we compute Word Error Rate (WER) and Character Error Rate (CER) for three systems: ElevenLabs Scribe v1, faster-whisper large-v3, and a fine-tuned Whisper small model trained on the same dataset's training split. ElevenLabs achieves a mean WER of 0.592 and mean CER of 0.185, outperforming faster-whisper large-v3 (WER 0.871, CER 0.266) on 44 of 50 clips. The fine-tuned Whisper small model achieves WER 0.368 on the validation set after 500 training steps – a 57.8% relative reduction versus base Whisper large-v3 and a 37.8% relative reduction versus ElevenLabs. **The fine-tuned result is in-domain and should not be directly compared to the commercial systems' out-of-domain performance; it demonstrates that fine-tuning on Somali data is highly effective and motivates future out-of-domain evaluation on broadcast audio.**

1 Introduction

SASR-Eval v0 (June 15, 2026) established that commercial ASR systems exhibit catastrophic failure modes on Djiboutian Somali speech: AssemblyAI Universal-2 outputs Arabic-dominant text in 100% of clips, and Whisper large-v3 produces trailing repetition loops on religious lecture audio. Those findings were heuristic – no ground-truth transcripts existed.

SASR-Eval v1 addresses this gap. Using the publicly available `skydheere/soomali-asr-dataset` (21,456 training / 2,146 validation / 2,146 test examples), we conduct the first quantitative WER/CER evaluation of commercial Somali ASR. We additionally train a fine-tuned Whisper small model on the dataset's training split and report in-domain validation performance, establishing a baseline for what targeted fine-tuning achieves before applying the approach to broadcast-domain data (the primary research target: Radio Télévision de Djibouti's 65-year archive and Radio Hargeysa's Somaliland corpus).

1.1 Relationship to SASR-Eval v0

AssemblyAI Universal-2 is excluded from v1 evaluation. SASR-Eval v0 established a 100% cross-lingual misroute rate (Arabic output on all 11 Somali clips); re-evaluation on a different dataset is unlikely to reverse a systematic routing failure. Resources were directed toward the fine-tuning experiment instead.

Table 1: Comparison of SASR-Eval v0 and v1 methodology

Dimension	v0	v1
Ground truth	None	HuggingFace dataset transcriptions
Clips	11 (60s each, 2 YouTube videos)	50 (10s each, HF test split)
Speakers	∼2 (BBC news, religious lecture)	Many (crowdsourced)
Metric	Heuristic (UWR, script detection)	WER, CER
Script mismatches	Present (AssemblyAI → Arabic)	None for either system
Fine-tuned model	Not evaluated	Whisper small, 500 steps, T4 GPU
AssemblyAI	Evaluated (failed)	Not re-evaluated (v0 failure conclusive)

2 Methodology

2.1 Dataset

The `skydheere/soomali-asr-dataset` is a publicly available Somali ASR dataset hosted on HuggingFace. It contains audio clips paired with human transcriptions in Somali Latin script. The test split used for commercial system evaluation contains 2,146 examples.

Dataset limitations relevant to interpretation:

- Speaker profile is likely crowdsourced read speech, not broadcast or spontaneous conversation
- Annotation artifacts are present in 34% of clips (pipe characters `|`, syntactic labels `OSV`/`SOV`/`SOVAV` leaked into reference transcriptions)
- Measured WER impact of artifacts after cleaning: ±0.002 (effectively zero; see Section 4.4)
- Somali dialect represented is likely standard/written Somali; Djiboutian broadcast dialect may differ

2.2 Clip Construction

50 clips were constructed from the HuggingFace test split as follows:

1. Sub-clips from the test split were concatenated with 200ms silence padding between examples
2. Each stitched clip targets approximately 10 seconds of audio
3. Ground-truth references were concatenated with space separation to match the stitched audio
4. Clips were exported as 16kHz mono WAV files

Stitching artifact: Adjacent sub-clips come from different speakers and contexts. Abrupt topic and speaker changes within a single clip are unnatural for production ASR systems designed for continuous single-speaker audio. Results on natural continuous broadcast speech may differ substantially.

2.3 ASR Systems Evaluated

Table 2 summarizes the three systems evaluated in this benchmark.

Table 2: ASR systems and configuration

Vendor	Model	Language hint	Notes
ElevenLabs	Scribe v1	`som`	Commercial API
faster-whisper	large-v3	`so`	Local inference, beam_size=5
Fine-tuned Whisper	small (ours)	`Somali`	Trained on HF train split, 500 steps

2.4 Fine-Tuned Model Training

The fine-tuned model was trained using the HuggingFace `transformers` Seq2SeqTrainer on the `skydheere/soomali-asr-dataset` training split (17,164 examples). Training configuration:

- **Base model:** `openai/whisper-small` (244M parameters)
- **Hardware:** Google Colab T4 GPU (free tier)
- **Training time:** ∼1 hour
- **Steps:** 500 (approximately 0.47 epochs)
- **Learning rate:** 1×10^{-5} with 100 warmup steps
- **Batch size:** 16 per device, no gradient accumulation
- **Evaluation:** Every 250 steps on the validation split (2,146 examples)

Critical caveat on comparability: The fine-tuned model was trained on the same dataset distribution as the evaluation. Its reported WER (0.368) is *in-domain performance* – the model has seen training examples from the same distribution as the test set. ElevenLabs and faster-whisper are evaluated *out-of-domain*: they were not trained on this dataset. Direct numerical comparison overstates the fine-tuned model's advantage. The appropriate interpretation is: (a) fine-tuning on Somali data is highly effective; (b) the gap between commercial systems and a fine-tuned specialist will likely persist on broadcast audio where neither has a training advantage; (c) out-of-domain evaluation on RTD broadcast clips is the correct next step.

2.5 Scoring

WER and CER were computed using the `jiwer` library. Text normalization applied before scoring:

```
def normalize(text):
    text = unicodedata.normalize("NFKC", text)
    text = re.sub(r"[’’‘´]", "’", text)   # apostrophe normalization
    text = text.lower()
    text = re.sub(r"[^\w\s’]", " ", text, flags=re.UNICODE)
    return re.sub(r"\s+", " ", text).strip()
```

Known latent bug (patched in v1): The v0 scorer treated Unicode curly apostrophes (`U+2019`) as punctuation and stripped them, while ASCII apostrophes (`U+0027`) were preserved. The v1 scorer normalizes all apostrophe variants to ASCII before lowercasing. No impact on this dataset (zero apostrophes in reference transcriptions), but the fix is necessary for other Somali corpora using apostrophes for pharyngeal sounds.

3 Results

3.1 Summary

Table 3: Summary results across 50 clips

System	n	Mean WER	Median WER	Mean CER	Median CER	Script Mismatches
Fine-tuned Whisper small (ours)	val	**0.368**	–	–	–	0
ElevenLabs Scribe v1	50	0.592	0.580	**0.185**	**0.167**	0
faster-whisper large-v3	50	0.871	0.862	0.266	0.253	0

Fine-tuned result is in-domain validation WER (not comparable to commercial out-of-domain results).

ElevenLabs Scribe v1 outperforms faster-whisper large-v3 on every metric across the 50-clip test set. The fine-tuned Whisper small achieves lower WER than both commercial systems on the validation split, though this comparison is confounded by domain overlap between training and evaluation data.

3.2 Commercial System Detailed Statistics

Table 4: ElevenLabs Scribe v1 – per-metric statistics (50 clips)

Metric	Mean	Median	Min	Max	Std Dev
WER	0.592	0.580	0.125	1.000	0.188
CER	0.185	0.167	0.020	0.613	0.105

Table 5: faster-whisper large-v3 – per-metric statistics (50 clips)

Metric	Mean	Median	Min	Max	Std Dev
WER	0.871	0.862	0.619	1.438	0.140
CER	0.266	0.253	0.136	0.433	0.063

Key observation: ElevenLabs has higher WER variance (SD 0.188 vs. 0.140), meaning higher upside (min WER 0.125) but also more variability. faster-whisper is consistently mediocre – its floor (min WER 0.619) is higher than ElevenLabs' floor, and its ceiling (max WER 1.438) is higher as well. ElevenLabs' lower CER (0.185 vs. 0.266) indicates it gets more characters correct even when word boundaries differ – suggesting better phonetic modeling of Somali.

3.3 Fine-Tuned Model Training Curve

Table 6: Fine-tuned Whisper small – training checkpoints

Step	Training Loss	Validation Loss	Validation WER	Notes
0	–	–	∼0.87	Base model (untrained)
250	0.447	0.460	0.513	Midpoint checkpoint
500	0.266	0.308	**0.368**	Final checkpoint

Training loss and validation loss remain closely tracking at step 500 (0.266 vs. 0.308), indicating the model is not overfitting at this stage. Continued training is likely to yield further

WER reduction. The WER improvement from step 0 to step 500 represents a 57.8% relative reduction versus the untrained Whisper large-v3 baseline on the same data distribution.

3.4 Head-to-Head: Commercial Systems Clip-by-Clip

ElevenLabs wins **44 of 50 clips** by WER. faster-whisper wins 4. 2 ties.

3.5 Notable Clips

3.5.1 Best ElevenLabs performance – clip50 (WER 0.125)

```
REF:        maalaa Guriga dadka soo dhisay ma aragtay? Waxaa yimid arday badan.
            Waxaa aan jeclaa Emma. gurigaygan
ElevenLabs: Maalaa guriga dadka soo dhisay ma aragtay. Waxaay yimid arday badan
            waxa aan jeclaa emma gurigaygan.
Whisper:    maa laa guriga dadka saadisay maa raqday waha yimid ardaybadan waha
            ancha laa emma guriga igan
```

ElevenLabs near-perfectly transcribes a full sentence. Whisper mangles several words ("saadisay" for "soo dhisay", "raqday" for "aragtay").

3.5.2 Best ElevenLabs performance – clip36 (WER 0.263)

```
REF:        Hargeysa waa magaalo weyn. Kani waa kaanaga. Warsame! farasey!
            Waan aqaannaa ardayda oo idil. maxkamadda gobolka qorayaasha iyo boqorrada
ElevenLabs: Hargeysa waa magaalo weyn. Kani waa kaannada. Warsame. Farasay.
            Waan aqaana ardayda oo idil. Maxkamadda gobolka. Qorshaha iyo boqorada.
Whisper:    Har geysa waa magaalaa weyn. Kani waa kaan naga. Huwar samay. Farasay.
            Waan aqaanaa radaayda oo eedil. Mahkamadda goblka. Urayaasha iyo baxar rada.
```

ElevenLabs preserves proper nouns ("Hargeysa", "Warsame", "Maxkamadda") and sentence structure. Whisper splits tokens incorrectly and garbles proper nouns.

3.5.3 Worst ElevenLabs performance – clip11 (WER 1.000)

```
REF:        goror caanaha saca raggaa Saddexda wiil ayaad ammaantay.
            OSV Dhiso. maalmo afar ah oo horreysay
ElevenLabs: Wuxuu u geystay qorayn ayaa ah goorar. Canaan saac. Ragga.
            Saddex daweyl ayaa dambe. Dhiso. Maalmo afar ah oo horaysay.
Whisper:    qorar aanaha saad ragga sada dhawiyil aya dhammaante diso
            maalmo qafar ah oo horay say
```

An annotation artifact ("OSV") in the reference and ambiguous audio caused ElevenLabs to hallucinate structure. Whisper echoed more of the phonetic surface form and achieved lower WER on this clip.

3.5.4 Largest gap – clip28 (WER: Whisper 1.438 vs. ElevenLabs 0.875)

```
REF:        qalimmada silsiladahayga dukaanle Ninkii libaaxii qabtay miyuu la hadlay?
            S/SOVAV baaldiyo labo ganacsato oo waawanaagsan
ElevenLabs: Xalammada silsiladaha iga dukaan leh ninkii libaaxi qabtay miyu la hadlay
            baal diyo laba ganacsi ugu wanaagsan.
Whisper:    Qalima dha. Silsila dhaha yiga. Du kaan la. Ninkin li baax hii qabtay.
            Miuu lahad la. Baal diyo. La baga naa sato'u buwan-buwan ahaksan.
```

Table 7: Per-clip WER and CER for both commercial systems

Clip	WER (Whisper)	WER (ElevenLabs)	CER (Whisper)	CER (ElevenLabs)	Winner
clip01	0.938	0.438	0.240	0.104	ElevenLabs
clip02	0.684	0.526	0.189	0.117	ElevenLabs
clip03	0.812	0.688	0.253	0.190	ElevenLabs
clip04	0.722	0.556	0.182	0.111	ElevenLabs
clip05	0.957	0.696	0.345	0.172	ElevenLabs
clip06	0.850	0.650	0.315	0.222	ElevenLabs
clip07	0.850	0.650	0.242	0.188	ElevenLabs
clip08	0.727	0.818	0.177	0.165	**Whisper**
clip09	0.882	0.588	0.288	0.154	ElevenLabs
clip10	0.905	0.857	0.331	0.297	ElevenLabs
clip11	0.867	1.000	0.269	0.613	**Whisper**
clip12	0.923	0.846	0.315	0.191	ElevenLabs
clip13	0.857	0.762	0.421	0.331	ElevenLabs
clip14	0.667	0.917	0.235	0.185	**Whisper**
clip15	0.900	0.500	0.267	0.171	ElevenLabs
clip16	0.824	0.529	0.183	0.128	ElevenLabs
clip17	0.824	0.647	0.290	0.370	ElevenLabs
clip18	0.667	0.500	0.244	0.078	ElevenLabs
clip19	1.231	0.462	0.253	0.165	ElevenLabs
clip20	0.917	0.500	0.329	0.139	ElevenLabs
clip21	0.809	0.714	0.344	0.264	ElevenLabs
clip22	0.842	0.737	0.245	0.431	ElevenLabs
clip23	0.765	0.882	0.289	0.333	**Whisper**
clip24	0.944	0.667	0.336	0.195	ElevenLabs
clip25	1.000	0.625	0.302	0.170	ElevenLabs
clip26	0.765	0.412	0.247	0.129	ElevenLabs
clip27	0.950	0.700	0.337	0.286	ElevenLabs
clip28	1.438	0.875	0.367	0.225	ElevenLabs
clip29	0.818	0.682	0.234	0.175	ElevenLabs
clip30	0.867	0.400	0.250	0.115	ElevenLabs
clip31	0.733	0.333	0.202	0.106	ElevenLabs
clip32	0.889	0.556	0.246	0.119	ElevenLabs
clip33	0.950	0.400	0.225	0.100	ElevenLabs
clip34	0.938	0.562	0.281	0.123	ElevenLabs
clip35	0.769	0.308	0.136	0.091	ElevenLabs
clip36	0.842	0.263	0.189	0.076	ElevenLabs
clip37	0.778	0.333	0.245	0.098	ElevenLabs
clip38	1.000	0.818	0.337	0.157	ElevenLabs
clip39	0.941	0.529	0.216	0.171	ElevenLabs
clip40	0.682	0.682	0.309	0.245	Tie
clip41	0.800	0.267	0.217	0.066	ElevenLabs
clip42	0.833	0.417	0.253	0.203	ElevenLabs
clip43	1.059	0.529	0.233	0.093	ElevenLabs
clip44	1.000	0.571	0.433	0.156	ElevenLabs
clip45	0.923	0.538	0.298	0.149	ElevenLabs
clip46	1.000	0.667	0.318	0.209	ElevenLabs
clip47	0.933	0.600	0.253	0.200	ElevenLabs
clip48	0.789	0.789	0.216	0.351	Tie
clip49	0.619	0.476	0.169	0.113	ElevenLabs
clip50	0.875	0.125	0.214	0.020	ElevenLabs

Whisper fragments continuous speech into spurious sentence breaks, inflating token count and pushing WER above 1.0. ElevenLabs handles run-on sentence structure substantially better.

4 Discussion

4.1 Interpretation of WER Values

The aggregate WER figures (0.59–0.87) are high in absolute terms but partially reflect the evaluation setup rather than ASR quality alone:

1. **Short reference clips inflate WER.** Many source clips are single words or short phrases. A single substitution on a 1-word reference yields WER = 1.0; an insertion yields WER > 1.0. CER is a more stable metric for this distribution.

2. **Stitching discontinuity.** Adjacent clips from different speakers with 200ms silence create conditions that production ASR systems are not designed for.

3. **Annotation noise.** 17 of 50 references (34%) contain artifacts; measured WER impact is ± 0.002 (see Section 4.4).

4. **CER tells a cleaner story.** ElevenLabs' mean CER of 0.185 indicates $\sim$81.5% character-level accuracy – consistent with usable quality for a low-resource language evaluation.

4.2 What Fine-Tuning Demonstrates

The fine-tuned Whisper small model achieves WER 0.368 in-domain – a 37.8% relative reduction versus ElevenLabs (0.592) and 57.8% versus base Whisper large-v3 (0.871) on the same data distribution. This in-domain result should not be interpreted as evidence that the fine-tuned model would outperform ElevenLabs on unseen broadcast audio.

The appropriate conclusion is narrower and more important: **a smaller model (Whisper small, 244M parameters) fine-tuned for approximately one hour on a single T4 GPU achieves substantially lower WER than a large commercial model (faster-whisper large-v3) on Somali data.** This demonstrates that the performance gap between commercial systems and a domain-specialized model is large and closeable with targeted training data.

The RTD archive and Radio Hargeysa corpus represent the out-of-domain broadcast data needed to validate this hypothesis on the actual target domain. SASR-Eval v2 will evaluate all three systems on broadcast clips with native-speaker ground truth.

4.3 Commercial System Comparison

ElevenLabs Scribe v1 is the stronger commercial system for Somali. It outperforms faster-whisper large-v3 on 44 of 50 clips by WER and achieves substantially lower CER (0.185 vs. 0.266). The WER standard deviation for ElevenLabs (0.188) is higher than Whisper (0.140), indicating ElevenLabs has more variability – it can achieve near-perfect transcription on some clips (WER 0.125) while failing completely on others (WER 1.000 on clip11). Whisper is more consistently mediocre across the full range.

For production applications requiring reliable Somali transcription on crowdsourced read speech, ElevenLabs is the preferred commercial option. Neither system has been validated on broadcast or dialectal Somali.

Table 8: Annotation artifact types and clip counts

Artifact type	Clips affected	Description
`|` (pipe)	14	Word-boundary or pause marker used by annotators
`OSV`, `SOV`, `SVO`	5	Syntactic word-order labels leaked into transcriptions
`SOVAV`	1	Extended annotation label

Table 9: WER impact of artifact removal

System	WER (original)	WER (cleaned refs)	Delta
ElevenLabs Scribe v1	0.592	0.591	−0.001
faster-whisper large-v3	0.871	0.873	+0.002

4.4 Annotation Artifact Analysis

17 of 50 clips (34%) contain annotation artifacts in the ground-truth reference:

The aggregate effect of artifact removal is effectively zero ($\leq \pm 0.002$ WER). Pipe characters were already handled by the punctuation regex; syntactic label tokens represent one word in an ∼18-word reference, yielding ∼5% denominator change. On individual clips, cleaning can increase WER (clip28 Whisper: 1.438 → 1.533) when a reference artifact token happened to absorb a deletion penalty.

5 Limitations

- **In-domain confound for fine-tuned model.** The fine-tuned Whisper small was trained and evaluated on the same dataset family. WER 0.368 is not directly comparable to commercial system WER on held-out data.
- **Stitching artifacts.** 200ms silence between unrelated speakers is unnatural for production ASR. Results on natural continuous speech will differ.
- **Short reference WER inflation.** Single-word references yield WER ≥ 1.0 for any single substitution. CER is the more interpretable primary metric.
- **Ground truth not re-verified.** The `skydheere/soomali-asr-dataset` transcriptions were taken as-is. Native-speaker validation has not been performed.
- **Single run per (clip, system).** No variance estimate. All systems are deterministic given fixed inference parameters.
- **No broadcast domain evaluation.** The dataset represents crowdsourced read speech. Generalization to RTD broadcast audio (the primary research target) is not established and is the subject of SASR-Eval v2.
- **Djiboutian dialect not represented.** The dataset likely represents standard written Somali. Djiboutian Somali (target dialect for RTD evaluation) may differ substantially.

6 Next Steps: SASR-Eval v2

SASR-Eval v2 will address the primary limitation of both v0 and v1: the absence of ground-truth evaluation on actual broadcast audio. Planned methodology:

1. **Audio collection.** RTD (Radio Télévision de Djibouti) and Radio Hargeysa broadcast clips, extracted from archive recordings during August 2026 fieldwork.

2. **Ground truth.** Native Somali and Afar speaker transcriptions with dialect annotation (Djiboutian Somali, Somaliland Somali, Afar).

3. **Systems evaluated.** ElevenLabs Scribe v1 (strongest commercial baseline), fine-tuned Whisper small trained on broadcast data (model retrained on RTD-derived corpus), and Afar ASR baseline (first evaluation of any ASR system on Afar broadcast audio).

4. **Evaluation design.** Natural continuous broadcast clips (60–120s), multiple speakers, news and cultural programming domains. No stitching artifacts.

SASR-Eval v2 will constitute the first rigorous out-of-domain evaluation of Somali ASR on Djiboutian broadcast speech and the first ASR evaluation on Afar in any domain.

7 Conclusion

SASR-Eval v1 establishes the first ground-truth WER/CER evaluation of commercial ASR systems on Somali speech. ElevenLabs Scribe v1 (WER 0.592) substantially outperforms faster-whisper large-v3 (WER 0.871) across 50 clips, winning 44 of 50 head-to-head comparisons. A fine-tuned Whisper small model trained for one hour on a single T4 GPU achieves WER 0.368 in-domain – demonstrating that targeted fine-tuning closes the gap between commercial general-purpose systems and domain-specialized models, and motivating the collection of broadcast-domain training data from the RTD and Radio Hargeysa archives.

The SASR-Eval series establishes a reproducible evaluation framework for Somali ASR. Native Somali and Afar speakers interested in contributing ground-truth transcriptions for SASR-Eval v2 should contact the authors.

Acknowledgments

The authors thank the maintainers of the open-source faster-whisper project, the `jiwer` library, and the HuggingFace `transformers` and `datasets` libraries. The `skydheere/soomali-asr-dataset` is the only publicly available ground-truth Somali ASR dataset known to the authors at time of writing. Audio processing used `yt-dlp`, `ffmpeg`, and `soundfile`.

References

[1] ElevenLabs. (2025). Scribe v1 API documentation. https://elevenlabs.io

[2] Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust Speech Recognition via Large-Scale Weak Supervision. *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

[3] Bain, M. (2023). faster-whisper: Reimplementation of OpenAI's Whisper model using CTranslate2. https://github.com/SYSTRAN/faster-whisper

[4] skydheere. (2024). soomali-asr-dataset. HuggingFace Datasets. https://huggingface.co/datasets/skydheere/soomali-asr-dataset

[5] Morris, A. (2020). jiwer: Evaluate your speech-to-text system. https://github.com/jitsi/jiwer

[6] Wolf, T., et al. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of EMNLP 2020: System Demonstrations*, 38–45.

[7] SASR-Eval v0 Report. (2026, June 15). Somali ASR System Evaluation: A Reference-Free Analysis of Production Models on Djiboutian Broadcast and Religious Speech.